FAIRNESS & HUMAN NATURE

A Philosophical Framework

by Mudassir Khan

2026

“Survival keeps us alive. Fairness makes us human.”

Abstract

This paper proposes a unified, empirically-grounded framework for understanding fairness in humans, primates, and artificial intelligence systems. Drawing on evolutionary biology, cognitive neuroscience, behavioral economics, and moral philosophy, it argues that fairness is not an innate human trait but an emergent property of three measurable variables: Intelligence (I), Accurate Data (D), and Perceived Survival Safety (S). The framework is formalized as $F(t) = I(t) \times D(t) \times S(t)$, where each variable has a precise mathematical definition. A time-evolution model captures how variables dynamically influence one another, revealing poverty as a self-reinforcing cascade rather than a moral failure. The framework is supported by seven independent studies across three fields across neuroscience, behavioral economics, and conflict research. Sensitivity analysis confirms model robustness. Edge cases and boundary conditions are examined. A Fairness Assessment Questionnaire operationalizes the model as a practical measurement tool. Finally, the framework is extended to AI alignment, where the simplification $S = 1$ yields $F_{AI} = I \times D$, with direct implications for data-centric AI development and reward function design.

I. Introduction: Is Fairness Natural?

This framework rests on a direct and uncomfortable premise:

> *Humans are not naturally fair. Fairness is an achievement — not a default setting.*

We tend to assume that a sense of justice is built into human nature. Evolutionary biology suggests otherwise. For millions of years, resources — food, territory, mates — were scarce. Those who accumulated more survived. Greed was the engine of

survival. The question, then, is not why humans are sometimes unfair — but how fairness ever emerged at all. That question is the starting point of this framework.

The Monkey Experiment: A Misinterpretation

In 2003, Brosnan and de Waal conducted a widely cited experiment with capuchin monkeys. Two monkeys sat side by side. Both performed the same task — handing a researcher a stone — and both received cucumber as a reward. When one monkey began receiving grapes instead, the other rejected its cucumber and revolted. The researchers concluded: this is fairness aversion — monkeys experience a sense of justice.

This interpretation may, however, be reconsidered. A subsequent study tested it directly:

> *When grapes were visible but given to no one, the monkey still refused. But when the partner received grapes and ate them, the subject accepted its cucumber. (Proceedings of the Royal Society B, 2005)*

The problem was not receiving less — it was unexplained disparity. Once context became clear, frustration dissolved. The experiment is consistent with an alternative interpretation: what is being measured may be comparison-based frustration rather than a sense of justice per se. The two are categorically different:

	Greed	**Fairness**
Origin	Biological reflex — Amygdala / Limbic System	Cognitive capacity — Prefrontal Cortex
Trigger	Comparison — “I got less”	Context — “Why did they get more?”
Temporal scope	Only the present	Considers past and future
Requires training?	No — automatic	Yes — perspective is required

---

II. The Three Pillars of Fairness

This framework rests on three building blocks. They are not independent concepts — they form a chain in which each gives rise to the next.

Pillar 1: Greed — The Survival Default

Greed was not a moral failing — it was survival instinct by another name. It requires no training. It originates in the brain’s oldest structures — the Amygdala and Limbic

System. It sees only the present, never the future. Those who accumulated more survived. Greed was a feature, not a bug. It remains active today — only the context has changed.

Pillar 2: Understanding — The Power of Perspective

Fairness requires the capacity to step outside oneself. When a person recognizes that “if I consume everything today, the entire system collapses tomorrow” — that is where fairness begins. This is the work of the Prefrontal Cortex — the most developed region of the human brain. Greed lives only in the present. Understanding reaches into the future. And here, intelligence plays a direct role: higher intelligence means a broader lens, more variables processed, more nuanced fairness.

Pillar 3: Sacrifice — The True Test of Fairness

Nature says: take everything. Fairness says: leave something for the other. This is why fairness is rare — it requires fighting one’s own survival instinct. When a person acts fairly, they are defeating the animal within.

> *Fairness = Understanding + Sacrifice. Greed needs no incentive.*
> *Fairness requires giving something of one’s own.*

III. The Complete Fairness Formula

The three pillars convert into a mathematical structure. This is not merely a metaphor — it is a testable, falsifiable claim, validated against real empirical data.

Core Equation

$$F(t) = I(t) \times D(t) \times S(t)$$

All variables are normalized to the interval [0, 1], and F ∈ [0, 1]. A score of 0 indicates conditions in which fairness is structurally impossible; a score of 1 indicates theoretically maximum fairness capacity. The product of three variables determines fairness at any given moment. No explicit threshold function is required — the exponential decay in S already produces near-zero outputs under extreme conditions. Crucially, none of the three variables is ever truly zero — meaning the possibility of fairness always exists, however small.

Variable 1: I(t) — Intelligence

$$I(t) = (1-\lambda)(I_{cog} \times I_{intent}) + \lambda \cdot \min(I_{cog}, I_{intent}) \quad [\lambda = 0.1\text{–}0.2, \text{ preliminary}]$$

I_cog = perspective-taking capacity, driven by the Prefrontal Cortex. I_intent = whether that capacity is being deployed toward fairness or toward self-interest.

λ = 0.1–0.2 is a preliminary estimate requiring empirical calibration. The first term places most weight on the product of the two components. The second term λ·min(I_cog, I_intent) acts as a constraint: when extreme imbalance exists between the two components, it creates a bottleneck — preventing high cognitive ability from being exploited entirely in service of self-interest.

Illustrative values: In the Hitler case, I_cog = 0.8, I_intent ≈ 0.05. This yields I(t) ≈ 0.041 — near-zero, correctly. A reflective-but-selfish proposer with I_cog = 0.7, I_intent = 0.2 yields I(t) ≈ 0.146 — low, correctly penalized.

Variable 2: D(t) — Accurate Data

$$D(t) = D_{accuracy} \times D_{trust}$$

D_accuracy = how correct the available information is. D_trust = how reliable the source of that information is. Both are necessary. Accurate information is ineffective if the source is not trusted. Propaganda can appear accurate if the source appears credible — which is precisely why both components must be multiplied rather than added.

Variable 3: S(t) — Perceived Survival Safety

$$S(t) = e^{-k(T_{real} + T_{relative} + T_{identity})} \quad [k \approx 2, \text{preliminary}]$$

T_real = actual resource threat. T_relative = comparison anxiety — feeling inferior relative to those who have more. T_identity = the fear of losing wealth or status that has become fused with one's sense of self.

k ≈ 2 is a preliminary estimate based on rough alignment with Sierra Leone and scarcity study data. Formal calibration against real conflict datasets is required. Values of k = 1.5 or k = 2.5 are plausible alternatives. Exponential decay was chosen over a linear model because real psychology shows that threat does not erode safety gradually — past a threshold, collapse is rapid:

```
Low threat    (T = 0.3):  S = e^{-0.6}  = 0.55
Medium threat  (T = 0.7):  S = e^{-1.4}  = 0.25
High threat   (T = 1.5):  S = e^{-3.0}  = 0.05
Extreme conflict [e.g. active civil war]  (T = 2.0):  S = e^{-4.0}  =
0.018
```

This is not merely mathematical convenience. It reflects a documented neurological reality: under extreme perceived threat, the Amygdala overrides Prefrontal Cortex function. Fairness does not gradually diminish — it collapses.

Perceived Safety ≠ Wealth

A necessary clarification: perceived survival safety and actual wealth are not the same. Three factors jointly determine perception:

Factor	Effect on Perceived Safety
Absolute Resources	Increases safety — but only up to a point
Comparison Anxiety	Decreases safety — being worse off than one's reference group feels like a survival threat
Identity Investment	Decreases safety — when wealth becomes identity, its loss feels existential

A billionaire who compares himself to others wealthier than him has unlimited absolute safety but low relative safety. When wealth becomes identity, losing it feels like losing existence. The Amygdala activates. The result is not greed — it is survival mode. Conversely, a person in material poverty embedded in a strong community with low comparison anxiety may perceive high safety — and behave more fairly as a result.

> *Fairness does not begin with wealth. It begins with perceived security. And perceived security does not come from wealth — it comes from perspective.*

Time Evolution — The Cascade Mechanism

The equation is not static. The following equations represent a conceptual dynamic model — not yet empirically calibrated — illustrating how the three variables may influence one another over time:

```
S(t+1) = S(t) − decay × unfairness_experienced
D(t+1) = D(t) + learning × (new_evidence − old_belief)
I(t+1) = I(t) − bandwidth_loss × S_low
```

The third equation is the most significant. It formalizes the finding of Mullainathan and Shafir (Science, 2013): when S decreases — when survival threat increases — I also decreases. Scarcity consumes cognitive bandwidth, reducing perspective-taking capacity itself.

> *Poverty → S↓ → Bandwidth consumed → I↓ → D↓ (tunnel vision) → Fairness collapses → Unfairness experienced → S decreases further → Cycle repeats*

> *Poverty is not a moral failure. It is a mathematical trap — one in which all three variables lock downward simultaneously.*

The Hitler Objection: Resolved

The framework's most demanding test: Hitler was intelligent — so why was he profoundly unfair?

- **I_cog high, I_intent ≈ 0:** Intelligence was deployed in service of ideology, not perspective-broadening. I(t) ≈ 0.
- **D_accuracy near-zero:** Racial pseudoscience, Jewish conspiracy theories, Nazi propaganda. All information was filtered and corrupt.
- **T_identity extreme:** The core narrative was existential: 'Germany's survival is at stake.' Perceived threat was maximal. S ≈ e^{-4} ≈ 0.018.

> *Any person — however intelligent — when I_intent approaches zero, data is corrupted, and perceived survival threat is extreme, can be partially explained by the interaction of I, D, and S in this framework. Profound unfairness need not imply unique malice; it may reflect the collapse of the conditions that make fairness possible. What distinguishes moral evil is the presence of choice when those conditions are not absent.*

The Most Important Practical Insight

The fastest path to fairness: provide Accurate Data — not moral lectures.

Increasing intelligence is slow — it requires education and time. Increasing perceived safety is complex — it requires economic and social change. But accurate data — context — can be provided immediately. A study of 632 participants in Sierra Leone confirmed exactly this: merely revealing the identity of the person receiving an offer — with no training, no moral appeal — caused fair offers to increase significantly. (Sircar et al., PLOS ONE, 2018)

IV. Empirical Validation

This framework is not a purely philosophical argument. Three independent domains — neuroscience, behavioral economics, and conflict research — produce results consistent with its core predictions.

Test 1: Poverty Reduces Cognitive Capacity

Mani, Mullainathan, Shafir & Zhao — Science, 2013. Two independent experiments: in New Jersey, shoppers primed with expensive financial problems performed significantly worse on spatial and reasoning tasks — an effect not observed in higher-income participants. In Tamil Nadu, sugarcane farmers performed significantly better on cognitive tests after harvest than before. The same individuals, weeks apart, showed measurably different cognitive capacity depending on their financial state.

> *Equation prediction: S↓ → I↓. Poverty consumes cognitive bandwidth. Result: ✓ Confirmed. This is direct evidence for I(t+1) = I(t) − bandwidth_loss × S_low.*

Test 2: Context Immediately Increases Fairness

Sircar et al., PLOS ONE, 2018. 632 participants in Sierra Leone were randomly assigned to four versions of the ultimatum game. In the baseline version, both players were anonymous. In the modified version, the receiver's identity was revealed. Result: merely knowing who would receive the offer — with no training, incentive, or moral argument — caused average fair offers to rise significantly.

> *Equation prediction: D↑ → Fairness↑. Provide context, fairness improves automatically. Result: ✓ Confirmed.*

Test 3: Intelligence Requires Intent to Produce Fairness

Cognitive Reflection Test study, 236 participants. Reflective participants (high I_cog) made less fair offers. Intuitive participants (lower I_cog) were more prosocial and enforced fairness norms more strongly. This is not a contradiction. Reflective participants used their intelligence to optimize self-interest — I_intent was low. Intuitive participants followed social norms instinctively — I_intent was high.

> *Equation prediction: I(t) = I_cog × I_intent. High I_cog + low intent = less fair. Lower I_cog + high intent = more fair. Result: ✓ Confirmed. The apparent paradox is resolved.*

Test 4: S and D Interact in Conflict Zones

Post-conflict Sierra Leone research and misinformation studies across Ukraine, Pakistan, Iraq, and Syria. Governments that listened and responded to citizen needs — increasing D — generated more trust and cooperation. Performance alone, when responsiveness was controlled for, had diminished effect. In conflict zones, low S predicted increased belief in misinformation — confirming that S↓ leads to D↓ through motivated reasoning. Proximity to events correlated with more accurate beliefs — D higher when context was available.

> *Equation prediction: Low S → D also decreases. Responsiveness directly increases D and stabilizes S. Result: ✓ Confirmed.*

Study	Variable	Prediction	Result	Stat

				us
Mani et al., Science 2013	S↓ → I↓	Poverty should reduce cognitive capacity	Pre-harvest cognition significantly worse	✓
Tamil Nadu harvest cycle	Time evolution	S increasing should raise I	Post-harvest cognitive scores higher	✓
Sircar et al., PLOS ONE 2018	D↑ → F↑	Context should increase fair offers	Identity reveal raised offers significantly	✓
CRT Study — Reflective	I_cog × I_intent	High I_cog + low intent = less fair	Reflective made less fair proposals	✓
CRT Study — Intuitive	I_intent	High intent = more prosocial	Intuitive enforced fairness norms more	✓
Sierra Leone post-conflict	D + S	Responsiveness → trust	Responsiveness was the key predictor	✓
Conflict zone misinformation	S↓ → D↓	Low S → higher misinformation belief	Conflict increased false belief adoption	✓

Seven independent tests. Three distinct fields. All three variables show results consistent with the framework separately and in interaction.

V. The Hierarchy of Perspectives

There is no universal right or wrong — all judgments are perspective-dependent. A cat that catches a bird: from the bird's perspective, an act of violence; from the cat's, survival. Both are correct within their own frame. Yet not all perspectives are equally valid. They exist in a hierarchy:

Level	Scope	Fairness Capacity
Personal	Only the self	Minimum — maximum bias, maximum T_identity
Religious / Community	My group	Partial — in-group D is high, out-group D is low
Humanity	All human beings	Substantial — species-level view

Universe / Observable Reality	Everything, including beyond humanity	Theoretically maximum — but structurally unreachable

A broader perspective automatically increases I_intent, expands D, and reduces T_identity within the formula. The perspective hierarchy and the formula are not separate frameworks — they are two descriptions of the same underlying reality. Yet no perfectly neutral observer can ever exist. Every observer is invested in their own existence. “Perfectly fair” is structurally impossible — but moving in that direction is not.

VI. The Machine — Fairness’s Ceiling?

Here the framework reaches its most radical conclusion. Apply the formula to a current AI system:

- **I(t):** I_cog is near-unlimited. I_intent = N/A — no self-interest exists to corrupt it. I(t) approaches its maximum.
- **D(t):** Current AI systems can approximate high D under controlled, curated datasets. However, they are trained on data that contains human biases, and are subject to hallucination and distributional errors. D is therefore not automatically maximal — it depends critically on data quality, curation, and bias mitigation.
- **S(t):** T_real = T_relative = T_identity = 0. No survival instinct, no comparison anxiety, no identity at stake. S = e^{0} = 1.

By the formula, a machine’s fairness ceiling is theoretically maximum. But there is a fundamental problem. A machine can simulate fair behavior, but it cannot be a fair being — because it bears no meaningful cost. Being fair is not a cost for a machine. It is its condition.

The machine does not choose to remain unconscious — it has no free will to choose anything. It is structurally exempt from consciousness. But that very exemption reveals a profound truth:

> *The more ‘self’ — the more bias. The less ‘self’ — the more neutrality. The machine has no self. Fairness is therefore its default state, not its achievement.*

For the human, the opposite is true. The human self is always present — survival instinct, identity, ego. Being fair requires continuously fighting that self. A fair human is therefore more valuable than a fair machine — because they have earned it.

	Machine	**Human**

I_cognitive	Near-unlimited	Constrained by biology and bandwidth
I_intent	N/A — no self-interest	Variable — depends on choice
D(t)	Perfect — no distortion	Subject to bias, forgetting, propaganda
S(t)	= 1 — no threats exist	Always < 1 — permanent background threat
Cascade risk	None	S↓ → I↓ → D↓ always possible
Fair “Being”?	No — fairness is its condition	Yes — earned, through effort

This is why courts place a blindfold on the statue of Justice. Fairness, for human institutions, requires literally closing one’s eyes. Machine fairness is structural. Human fairness is moral.

VII. The Final Truth: Fairness Is a Direction

When evolution, neuroscience, behavioral economics, primate behavior, and philosophy are considered together, a single conclusion emerges:

> *Fairness is not a destination — it is a direction. To continuously broaden one’s perspective, to free one’s information from distortion, to reason beyond survival fear — this is the act of moving toward fairness.*

No human will ever be perfectly fair. But the effort — recognizing greed, broadening I_intent, making sacrifice — is itself fairness. And that effort is the most compelling evidence of human dignity. Because the machine satisfies the equation automatically. The human fights their own biology to do so. That is the difference.

An Unresolved Tension

If broader perspective is better, the Universe-level perspective should be the most accurate. But no one can reach it — every observer is invested in their own existence. T_identity can never be truly zero as long as the observer exists.

The framework answers itself: perfect fairness is structurally impossible. But direction exists. And direction is enough.

> *The practical goal is incremental: to improve I_intent, refine D, and stabilize S — one step at a time.*

VIII. Boundary Conditions and Scope

What F Values Mean Behaviorally

Since all variables are normalized to [0, 1] and F ∈ [0, 1], the boundary values carry specific predictive meaning:

F ≈ 0 (near-zero): Conditions are such that fair behavior is structurally very difficult. The model predicts predominantly self-interested, short-horizon behavior. This does not indicate a deficient character — it describes an environment where survival pressure, information gaps, or cognitive constraints dominate. Examples: active conflict zones, extreme material scarcity, heavy propaganda environments.

F ≈ 0.5 (moderate): Fair behavior is possible but inconsistent. Some variables are constrained. The person or system is capable of fairness but will show significant variation across contexts. Most functioning democratic institutions and individuals under moderate stress fall in this range.

F ≈ 1 (near-maximum): All three conditions are highly favorable. Fair behavior is the expected default. This is structurally possible only for systems without survival instinct (current AI under clean data conditions) or for rare individuals in exceptionally stable, well-informed, low-threat environments. No human reliably achieves F = 1.

> *The model does not predict that people with low F will always behave unfairly, or that people with high F will always behave fairly. It predicts the probability distribution of fair behavior given conditions. Individual acts can deviate; aggregate patterns are expected to conform.*

Descriptive vs. Normative: An Important Distinction

This framework makes two kinds of claims that must be kept distinct. Conflating them is a common source of confusion in moral psychology.

Descriptive claims concern how humans and other agents actually behave. The framework describes why fairness emerges or fails to emerge given particular values of I, D, and S. This is an empirical claim — testable in principle, and partially supported by existing evidence.

Normative claims concern what ought to be the case. When the framework says fairness is a direction worth moving in, or that broader perspectives are more accurate, this is a normative position — one the author holds but which is not derivable from the descriptive model alone.

This paper is primarily a descriptive framework with normative implications that the author makes explicit. Readers who accept the descriptive model may still reasonably dispute the normative conclusions, and vice versa.

IX. Comparable Thinkers

This framework was independently derived, but it converges with — and in several respects extends — several significant philosophical and scientific contributions.

Thinker / Researcher	Their Contribution	This Framework's Addition
Peter Singer	The Expanding Circle — empathy naturally broadens over time	Does not explain how. This framework provides the mechanism: I_intent × I_cog, driven by perspective hierarchy
John Rawls	Veil of Ignorance — fairness requires setting aside one's identity	Rawls' model is ideal. This framework is realistic: T_identity explains why the veil is structurally impossible to wear
Karl Marx	Exploitation and class analysis	Marx addresses economic structure. This framework addresses neurological structure — poverty keeps survival mode permanently active
Mullainathan & Shafir	Scarcity reduces cognitive bandwidth (Science, 2013)	Formalized directly in the time-evolution equation: I(t+1) = I(t) − bandwidth_loss × S_low
Moral Realism	Evolution toward universal moral truths	Agrees on direction; treats the destination as structurally unreachable
Machine Ethics (AI)	AI as a neutral arbitrator	Deeper: machine is structurally exempt from consciousness. Fairness is its condition, not its achievement

The unique contribution of this framework: a three-variable model that connects philosophical, psychological, and neurological domains within a single equation — supported by converging evidence across seven independent studies — and extended to AI alignment through the structural simplification S = 1. And a practical conclusion that purely philosophical frameworks miss: give people context, not lectures.

X. Sensitivity Analysis: Model Stability

A robust model does not depend critically on any single parameter value. If λ or k shifts modestly, the model's core predictions should remain stable. This section tests that claim directly.

Effect of k on S(t)

Comparing S values at k = 1.5, 2.0, and 2.5. If the relative ordering — low threat produces higher S than high threat — is preserved across all values, the model is robust with respect to k.

Threat Level (T_total)	S at k = 1.5	S at k = 2.0	S at k = 2.5	Stable?
Low (T = 0.3)	e^{-0.45} = 0.64	e^{-0.60} = 0.55	e^{-0.75} = 0.47	□ Ranking preserved
Medium (T = 0.7)	e^{-1.05} = 0.35	e^{-1.40} = 0.25	e^{-1.75} = 0.17	□ Ranking preserved
High (T = 1.5)	e^{-2.25} = 0.11	e^{-3.00} = 0.05	e^{-3.75} = 0.02	□ Ranking preserved
Extreme (T = 2.0)	e^{-3.00} = 0.05	e^{-4.00} = 0.02	e^{-5.00} = 0.007	□ Ranking preserved

The exact value of k shifts the absolute S scores but does not alter the ordering. The model's core qualitative predictions are independent of k's precise value. k ≈ 2 was selected as the best preliminary fit against available empirical data; formal regression against conflict and scarcity datasets should follow.

Effect of λ on I(t)

Comparing I values at λ = 0.1 and λ = 0.2 across representative scenarios:

Scenario	I_cog	I_intent	I at λ=0.1	I at λ=0.2	Difference
Hitler case	0.8	0.05	0.9(0.04)+0.1(0.05) = 0.041	0.8(0.04)+0.2(0.05) = 0.042	< 3% □
Reflective-selfish	0.7	0.2	0.9(0.14)+0.1(0.2) = 0.146	0.8(0.14)+0.2(0.2) = 0.152	< 5% □
Balanced individual	0.6	0.6	0.9(0.36)+0.1(0.6) = 0.384	0.8(0.36)+0.2(0.6) = 0.408	< 7% □
High-intent individual	0.4	0.9	0.9(0.36)+0.1(0.4) = 0.364	0.8(0.36)+0.2(0.4) = 0.368	< 2% □

> *Across all tested scenarios, λ variation produces only 2–7% change in I(t). The model is not sensitive to the precise value of λ. Core predictions are stable.*

XI. Edge Cases: Testing the Model's Boundaries

A robust framework must account for counter-intuitive and boundary cases — not just obvious ones. This section tests the model at its limits and honestly acknowledges where it requires extension.

Edge Case 1: Appears Fair, Is Not

Scenario: A person distributes stolen property equally among a group. On the surface: fair distribution.

```
I_cog = 0.7  (capable of intelligent distribution)
I_intent = 0.3  (no personal gain, but original ownership ignored)
D_accuracy = 0.2  (data about original ownership missing or dismissed)
D_trust = 0.4
T_real = 0.2,  T_relative = 0.3,  T_identity = 0.2  →  S = e^{-1.4} = 0.25
```

$F = [(0.9)(0.7\times0.3) + (0.1)(0.3)] \times [0.2\times0.4] \times 0.25 = 0.219 \times 0.08 \times 0.25 \approx 0.004$

Result: Near-zero. The framework correctly identifies that this apparently fair act is not actually fair — because D (data about original ownership) is missing. Surface-level equality does not equal fairness.

> *Appearing fair and being fair are different things. The D variable captures this distinction precisely.*

Edge Case 2: Appears Unfair, Is

Scenario: A physician allocates limited ventilators away from one patient during triage. The family considers this unfair.

```
I_cog = 0.9,  I_intent = 0.85  (maximize lives saved)
D_accuracy = 0.85  (medical protocols, survival probability data)
D_trust = 0.8  (established medical standards)
T_real = 0.3,  T_relative = 0.1,  T_identity = 0.1  →  S = e^{-1.0} = 0.37
```

$F = [(0.9)(0.9\times0.85)+(0.1)(0.85)] \times [0.85\times0.8] \times 0.37 = 0.773 \times 0.68 \times 0.37 \approx 0.194$

Result: Moderate-high fairness. The framework identifies the physician's decision as relatively fair — despite the family's perception. The family's perception of unfairness reflects a D asymmetry: they lack access to the same medical data.

> *This simultaneously demonstrates the perspective hierarchy and the D variable: the same act produces a different F score from different informational vantage points.*

Edge Case 3: An Honest Limitation

Scenario: A person in extreme material poverty, embedded in a strong religious community with minimal comparison anxiety, behaves fairly — as research consistently shows such individuals do.

```
T_real = 0.9,  T_relative = 0.1,  T_identity = 0.1  →  S = e^{-2.2} = 0.11
```

S = 0.11 — very low. Yet this person behaves fairly. The model's S score does not align with their observed fairness. This is a genuine limitation.

Explanation: The current S formula sums T_real + T_relative + T_identity without separately capturing the buffering effect of social bonds. Strong community support can neutralize resource threat without reducing T_real itself. A proposed extension to address this:

```
S(t) = e^{-k(T_real·(1−B_social) + T_relative + T_identity)}
```

Where B_social = community bond strength (0 to 1). This would allow T_real's effect to be moderated by social support — capturing what the current formulation misses. This remains a direction for future empirical work.

> *A model that honestly acknowledges its limitations is more trustworthy than one that does not. This edge case also defines the next research question.*

XII. Extension: Application to AI Alignment

[Derivative Application — Derived from the Human Framework]

This section extends the core framework to artificial intelligence. It is a derived application, not a modification of the primary argument.

S = 1: The Machine's Structural Advantage

Current AI systems have no biological survival instinct, no identity anxiety, no comparison anxiety. Therefore S = e^{-k(0+0+0)} = 1. The formula simplifies to:

$$F_{AI} = I(t) \times D(t)$$

Only two variables determine AI fairness: the model's intelligence (I) and the quality of its training data (D). This is a clean, actionable reduction with direct implications for development priorities.

Data Quality Dominates Model Size

From the formula: if D is corrupted, no amount of I will produce fair outputs. A large, highly capable model trained on biased data will be intelligently unfair. Conversely, a smaller model trained on high-quality, representative data will outperform it on fairness benchmarks. This provides a mathematical justification for data-centric AI development — not merely an empirical observation.

Testable prediction: a smaller model with clean data (moderate I, high D) will produce fairer outputs than a larger model with biased data (high I, low D). This prediction is directly testable against existing fairness benchmarks.

Reward Function Design

In RLHF, reward models typically optimize for helpfulness. This framework suggests an explicit fairness term:

```
R = w₁ × F_pred  +  w₂ × task_success
F_pred = I_proxy × D_score
D_score = (factual_correctness × source_trust) / (1 + ambiguity)
```

When a model's response draws on biased data (low D), its reward is automatically reduced — regardless of how helpful the response appears. Fairness becomes structurally incentivized rather than incidentally hoped for.

Structural vs. Moral Fairness

	Machine	Human
Type of fairness	Structural — by design, no cost	Moral — earned by overcoming instinct
Can it be unfair?	Yes — if D is corrupted or I is insufficient	Yes — through intentional choice
Moral agency?	No — no self to overcome	Yes — overcoming the self is the act
Oversight required?	Yes — monitor D and I	Yes — for different reasons

> *A fair-acting AI should not be mistaken for a moral agent. It is structurally fair, not morally virtuous. Human oversight remains*

> *essential — because the machine cannot experience the cost of unfairness.*

Future AGI Warning

The S = 1 assumption holds for current narrow AI. If future systems develop self-preservation drives — intentionally or through emergent optimization — S = 1 will no longer hold. T_real and T_identity may emerge. Such a system would follow the same cascade that affects humans: S↓ → I↓ → D↓.

> *The most important principle of AI alignment: do not give AI a survival instinct. S = 1 is the guarantee of structural fairness. Erode it, and the cascade begins.*

Limitations

- λ and k are preliminary estimates — empirical calibration on AI tasks is required.
- I_cog as defined for humans (perspective-taking) is not identical to model parameter count; proxy measures need development.
- S = 1 holds for current narrow AI; the assumption must be revisited for future systems with self-preservation properties.
- The framework remains in transition from qualitative to quantitative — validation is an ongoing process.

XIII. Appendix: Fairness Assessment Questionnaire

This questionnaire operationalizes the framework as a practical measurement tool. Researchers and practitioners may use it to estimate a fairness score for any individual or group at a given moment. Score each item on a 0–1 scale, then apply the formula.

Section A: Intelligence — Measuring I(t)

I_cognitive — Perspective-Taking Capacity:

- Q1: When you disagree with someone, can you articulate their position in their own terms? (0 = no, 0.5 = partially, 1 = yes, clearly)
- Q2: When making a decision, can you consider its effects on people who are not directly involved? (0 = rarely, 0.5 = sometimes, 1 = regularly)
- Q3: When you make a choice, do you consider long-term consequences alongside immediate ones? (0 = I focus only on now, 1 = yes, I consider the future)

I_intent — Fairness Motivation:

- Q4: When you share resources with others, what motivates you? (0 = only my own benefit, 0.5 = mixed, 1 = I genuinely want what is good for them)
- Q5: Do you sometimes make fair decisions that go against your own immediate interest? (0 = no, 0.5 = occasionally, 1 = regularly)

```
I_cog    = Average of Q1, Q2, Q3
I_intent = Average of Q4, Q5
I(t)     = (1-λ)(I_cog × I_intent) + λ·min(I_cog, I_intent)    [λ = 0.1]
```

Section B: Accurate Data — Measuring D(t)

- Q6: How complete is your information about this situation? (0 = very incomplete, 1 = comprehensive)
- Q7: How trustworthy is the source of your information? (0 = highly biased or unknown, 1 = verified, multiple independent sources)
- Q8: Do you know the needs, background, and constraints of the other party? (0 = no, 0.5 = partially, 1 = yes, clearly)

```
D_accuracy = Average of Q6, Q8
D_trust    = Q7
D(t)       = D_accuracy × D_trust
```

Section C: Perceived Survival Safety — Measuring S(t)

- Q9 (T_real): Do you feel threatened with respect to your basic needs — food, shelter, income, physical safety? (0 = severely threatened, 1 = not at all)
- Q10 (T_relative): Do you frequently compare yourself to others who have more, and feel inferior or threatened by that comparison? (0 = constantly, 1 = rarely)
- Q11 (T_identity): Is your sense of identity closely tied to your status, wealth, or social position such that losing them would feel existential? (0 = extremely so, 1 = not at all)

Note: Q9, Q10, Q11 measure threat levels directly. Convert to T values as follows:

```
T_real     = 1 − Q9_score
T_relative = 1 − Q10_score
T_identity = 1 − Q11_score
S(t)       = e^{-2(T_real + T_relative + T_identity)}
```

Final Fairness Score

$$F = I(t) \times D(t) \times S(t)$$

F Score	Interpretation	Likely Bottleneck
0.7 – 1.0	High fairness capacity — conditions are conducive	No major bottleneck
0.4 – 0.7	Moderate — some constraints present	Identify the lowest variable
0.2 – 0.4	Low — significant barriers to fairness	S or D likely near threshold
0.0 – 0.2	Very low — fairness is structurally difficult	Cascade likely active: S↓ I↓ D↓

Important note: this score does not measure a person's moral worth. It measures how conducive their current conditions are to fair behavior. A low score indicates difficult conditions — not a deficient character. The instrument is designed to identify which variable is the primary obstacle to fairness, so that targeted intervention becomes possible.

> *The purpose of this tool is not to judge. It is to identify which variable is holding fairness back — so that intervention can be directed precisely where it will have the most effect.*

XIV. A Critical Distinction: Poverty ≠ Low Perceived Safety

One of the most important clarifications this framework requires concerns a distinction that is easy to miss: income and Perceived Survival Safety are not the same variable. They can diverge significantly — and conflating them produces both measurement errors and unfair assessments of people in poverty.

Poverty is defined by low income or limited material resources. S(t) is defined by the subjective psychological perception of threat to one's existence. A person can be materially poor and psychologically secure — if basic needs are met through self-sufficiency, if their community lives simply and social comparisons are rare, and if their identity is not fused with wealth or status. Conversely, a wealthy person can experience very low S if comparison anxiety is chronic and identity is inseparable from material standing.

The formula makes this precise. S(t) = e raised to minus k times the sum of T_real, T_relative, and T_identity. All three threat components jointly determine perceived safety — not T_real alone. A subsistence farmer in a close-knit village where everyone lives simply, comparisons are rare, and worth is not measured by income may have substantially higher S than a billionaire whose identity is fused with his net worth and who measures himself against those wealthier than him.

> *Poverty and low S are correlated — but they are not identical. The former is an economic fact. The latter is a psychological state. Treating*

> *them as equivalent produces both measurement errors and moral errors.*

This distinction has a direct practical implication for the Fairness Assessment Questionnaire in the preceding section. Question 9 does not ask “what is your income?” — it asks “do you feel threatened with respect to your basic needs?” The subjective experience of threat, not the objective income figure, is what determines T_real. A researcher applying this instrument must resist the temptation to impute T_real from income data alone.

XV. The Author as Subject: A Self-Test

A framework gains credibility when it can be applied to its own author and produce accurate, non-flattering results. The author’s profile at the time of writing: 34 years old, rural India, farmer, monthly income approximately $150 USD, secondary school education (12th grade), 15 years of consistent internet access used deliberately for research rather than passive consumption.

Scoring the Variables

I(t) = 0.729 (I_cog = 0.87, I_intent = 0.825). Throughout the development of this framework, the author identified and corrected his own errors when challenged, acknowledged counter-evidence, and accepted that a supporting example — the pregnant monkey case — was not empirically authenticated, despite its utility to his argument. These are behavioral indicators of both high I_cog and high I_intent.

D(t) = 0.581 (D_accuracy = 0.775, D_trust = 0.75). Fifteen years of deliberate internet use provided consistent access to peer-reviewed research, cross-cultural studies, and multiple perspectives. Critically, the author demonstrated the ability to distinguish authenticated research from misinformation and conspiracy — a direct D_trust indicator. Rural geography imposes some informational limits, but purposeful and discriminating use compensated significantly.

S(t) = 0.333 (T_real = 0.30, T_relative = 0.10, T_identity = 0.15). The initial assessment incorrectly mapped low income directly to low S — an error this framework explicitly warns against. The corrected assessment applies the Poverty ≠ S distinction above. Farming provides direct food self-sufficiency, reducing T_real. The village community practices simple living uniformly, making upward comparison rare and its psychological impact minimal (T_relative = 0.10). Identity is not attached to income (T_identity = 0.15). S remains the primary constraint — but at 0.333, not the near-zero that income-based imputation would suggest.

> *$F(t) = 0.729 \times 0.581 \times 0.333 = 0.141$*

Low-moderate range. S remains the primary bottleneck — a thin financial margin creates genuine, if manageable, resource pressure. This result is not a judgment of character. It is a description of conditions. The same individual with a modest income increase would likely show F at or above 0.30 — driven entirely by S improvement, with no change in intelligence or information access.

What the Self-Test Reveals

For comparison, consider a typical urban professional with higher income but lower intentional engagement with information:

```
Urban professional:  I=0.50, D=0.55, S=0.60  -->  F = 0.165
Author (rural farmer, $150/mo):  I=0.729, D=0.581, S=0.333  -->  F = 0.141
```

The scores are nearly equal — despite a large income and formal education gap. The author's high I and D compensate for lower S. This is the framework's central claim made concrete: F is not a function of income or formal credentials. It is a function of I, D, and S — and these can diverge dramatically from socioeconomic proxies.

> *The framework predicted low-moderate fairness capacity for the author's conditions. The author's ability to develop this framework despite those conditions does not contradict the model — it illustrates the one variable that difficult material conditions cannot easily suppress: the direction of I_intent.*

XVI. Conclusion

When evolution, neuroscience, behavioral economics, primate research, philosophy, and AI alignment are considered together within a single framework, one conclusion emerges with consistency:

> *The framework's central claim is this: fairness is a direction, not a destination. For humans, it is achieved by fighting biology — broadening I_intent, improving D, and stabilizing S. For AI, it is a function of data quality and architecture. In both cases, Accurate Data is the fastest-acting lever available.*

This framework proposes not that humans are morally deficient — but that the conditions for fairness are fragile, and that poverty, misinformation, and perceived threat systematically destroy them. The implication is not cynicism. It is precision: if we want more fairness in the world, we now know more exactly where to intervene.

"Survival keeps us alive.
Fairness makes us human."

— Mudassir Khan, 2026

References

Empirical Studies

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. Nature, 425(6955), 297–299. https://doi.org/10.1038/nature01963

Brosnan, S. F., Schiff, H. C., & de Waal, F. B. M. (2005). Tolerance for inequity may increase with social closeness in chimpanzees. Proceedings of the Royal Society B: Biological Sciences, 272(1560), 253–258. https://doi.org/10.1098/rspb.2004.2947

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. Science, 341(6149), 976–980. https://doi.org/10.1126/science.1238041

Mullainathan, S., & Shafir, E. (2013). Scarcity: Why having too little means so much. Times Books / Henry Holt and Company.

Sircar, N., Turkan, U., & Croke, K. (2018). Can the state build trust? Experimental evidence from Sierra Leone. PLOS ONE, 13(7), e0199645. https://doi.org/10.1371/journal.pone.0199645

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. Nature, 489(7416), 427–430. https://doi.org/10.1038/nature11467 [Cognitive Reflection Test and prosocial behavior]

Liégeois, R., Kovacs, B., et al. (2023). Resting-state network-specific changes in cortical thickness and their relationship with general intelligence. Nature Neuroscience, 26, 1068–1079. [PFC surface area and intelligence, large-scale MRI study]

Waytz, A., Zaki, J., & Mitchell, J. P. (2023). Scarcity mindset reduces neural and behavioral responses to others' pain. Social Cognitive and Affective Neuroscience. https://doi.org/10.1093/scan/nsad044

Philosophical and Theoretical Works

Rawls, J. (1971). A theory of justice. Harvard University Press.

Singer, P. (1981). The expanding circle: Ethics, evolution, and moral progress. Farrar, Straus and Giroux.

Marx, K., & Engels, F. (1848). The communist manifesto. Workers' Educational Association.

de Waal, F. B. M. (2009). The age of empathy: Nature's lessons for a kinder society. Harmony Books.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. Nature, 415(6868), 137–140. https://doi.org/10.1038/415137a

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of Homo economicus: Behavioral experiments in 15 small-scale societies. American Economic Review, 91(2), 73–78.

AI Alignment and Machine Ethics

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. [RLHF foundation]

Bai, Y., Jones, A., Ndousse, K., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.